

# A computer vision model for visual-object-based attention and eye movements



Sun, Y., Fisher, R., Wang, F., & Gomes, H. M. (2008). A computer vision model for visual-object-based attention and eye movements. Computer Vision and Image Understanding, 112(2), 126-142doi: 10.1016/j.cviu.2008.01.005

# A Computer Vision Model for Visual-Object-Based Attention and Eye Movements

Yaoru Sun [a], Robert Fisher [b], Fang Wang [c] and Herman Martins Gomes [d]

[a] *Nanjing University, China; Brain and Behavioural Sciences Centre, University of Birmingham, Birmingham B15 2TT, UK. Email: yaorus@gmail.com*

[b] *School of Informatics, University of Edinburgh, JCMB, The King's Buildings, Edinburgh EH9 3JZ, UK. Email: rbf@inf.ed.ac.uk*

[c] *The Intelligent Systems Lab, BT Exact, Ipswich, IP5 3RE, UK. Email: fang.wang@bt.com*

[d] *Universidade Federal de Campina Grande Departamento de Sistemas e Computação Av. Aprgio Veloso s/n 58109-970 Campina Grande PB Brazil. Email: hmg@dsc.ufcg.edu.br*

**Abstract**

This paper presents a new computational framework for modelling visual-object based attention and attention-driven eye movements within an integrated system in a biologically inspired approach. Attention operates at multiple levels of visual selection by space, feature, object and group depending on the nature of targets and visual tasks. Attentional shifts and gaze shifts are constructed upon their common process circuits and control mechanisms but also separated from their different function roles, working together to fulfil flexible visual selection tasks in complicated visual environments. The framework integrates the important aspects of human visual attention and eye movements resulting in sophisticated performance in complicated natural scenes. The proposed approach aims at exploring a useful visual selection system for computer vision, especially for usage in cluttered natural visual environments.

*Key words:* Visual-object based competition, space-based attention, object-based attention, group-based attention, foveated imaging, attention-driven eye movements.

## 1 Introduction

Human vision uses visual attention to scrutinize important information with high acuity and select information relevant to current visual tasks. When interesting objects in the visual periphery need to be explored, attention may employ an eye movement to shift gaze to them for more detailed analysis. In this way, human vision can use limited visual sources to effectively deal with complex visual selection tasks [21, p. 53, p. 80-88]. Visual (covert [1]) attention has advantages of speediness, accuracy, and maintenance of mental processing and can freely undertake visual selection without eye movements, but may also need eye movements to extend the selection and improve performance in large-scale visual environments by a series of gaze shifts over time [9]. Because eye movements or fixation shifts take time and result in significantly different foveated images, the visual system must integrate these partially overlapping images in a spatio-temporal context for unified and coherent visual selection.

Visual attention and eye movements have been studied and used individually or jointly in numerous computer vision models. Most of these attention models (e.g., [12]) have been developed from the traditional psychophysical accounts of space-based attention which suppose what attention selects are locations regardless of being occupied by targets or not or even nothing at all. This therefore may lead to exhaustive "point-by-point" searching to find a possible target. Search efficiency would be disastrous if a scene is cluttered or contains only a few meaningful objects. The computer vision models of space-based attention were reviewed in [30] and many other works. A recent study [5] also showed that the space-based models may produce inconsistent selections under image similarity transforms and could not select both positions and scales. Furthermore, hierarchical selectivity of attention for structured objects, groups of objects, or objects overlapped in the same place is difficult for space-based models but can be naturally tackled by the object-based attention accounts which hold attention actually prefer selecting objects and proto-objects [27] [2].

In recent years, inspired by psychophysical research on object-based attention, some pioneer works have started to develop computational models of non-spatial attention for computer vision. By extending the Integrated or Biased Competition account for object-based attention [6], Sun and Fisher [30] developed the first computational framework for object-based attention with integrating space-based attention. Hierarchical selectivity or multiple selection

[1] The terms of "covert" and "overt" are used here specifically to distinguish between (covert) attention (shifts) versus (overt) eye movements.

[2] "Proto-object" is deemed to be "pre-attentive object", "visual object" or cluster of features formed without attention and solves the binding problem for attention, according to [23], [24]

of attention by feature, location, object, and group was achieved through the extended concept of "grouping". Tsotsos et al. have refined the Selective Tuning Model for region-based selection and extended it to active visual search and object motion domains [37], [34]. Orabona et al. [20] proposed an object-based attention model by the "blob" generating proto-object selection to guide the grasping and recognition of objects in a humanoid robot. Based on bottom-up and top-down interaction via an Interactive Spiking Neural Network and face-like region segmentation, Lee et al. [18] introduced a non-location model of attention to detect and select a potential region that may contain a face. By using the feedback connections in the saliency computational hierarchy with a operation of expanding a salient location to a salient region, Walther et al. [36] extended Itti et al. work [12] from salient location to salient region based selection.

Visual attention mechanisms have also been broadly used in a great number of active vision models to guide fovea shifts [20]. Typically, log-polar or foveated imaging techniques have been employed to simulate retinal sensing in the models of eye movements [28]. Recently, the saliency-based mapping approach has also been incorporated in some active vision models for eye movements [10], [1]. Nevertheless, modelling attentional shifts and gaze shifts based on their shared control circuits in a coherent system by their biologically-plausible relationship have not yet been found in any of these works. Both kinds of these shifts working cooperatively not only makes the system biologically-plausible, but also facilitates the practical usage in computer vision [15]. To implement this, the system firstly needs to build the coherent architecture and engine to properly drive attentional shifts and gaze shifts. In addition, the system needs to consider: 1) How and when attention engages with and disengages from eye movements; 2) How attention shifts under foveal fixation; 3) How and when attention programs (or drives) an eye movement to the periphery and comes back; 4) How attention can work at multiple selectivity by features, locations, objects and groups, so that shifts of attention and shifts of gaze can work together to provide effective visual selection.

The closely linked relationship between visual attention and eye movements has been revealed by a growing body of psychophysical and neuropsychological studies but how attentional shifts and gaze shifts are precisely related still remains open. The premotor theory of attention posits that the shift of attention is a covert unexecuted saccade and is identical to saccade planning [26], [4]. However, many other recent findings suggest that attention can shift freely to targets without shifting or planning gaze, and attention not only affects eye movements but also is a precursor for saccade generation [19], [14], [32]. In this work we accept that attention and eye movements are strongly linked as supported by convergent evidence from many studies including the premotor theories of attention. But we reject the premotor theory's premise about that attentional shift and saccadic shift or planning are identical and

attention processes are the consequence of motor programming. Rather, we adopt the following suggestions:

- Visual attention gates the information processing in sensory brain regions and serves to bias objects' competition for representation in favour of a relevant location, feature, object, or group of objects at multiple levels from early sensory processing to higher level visual processing, with the eventual winner and all of its features selected for action as suggested by the "biased" or "integrated" competition theory of attention [6], [3];
- Recent studies have revealed that visual attention can work at multiple levels of visual selection or hierarchical selectivity. Feature, location, object, or group based attention are not mutually exclusive. They may reflect different selection behaviour of a unitary attention mechanism depending on the nature of targets and visual tasks, and share many similar underlying attentional circuits and anatomical areas in an extensive network of brain regions [27], [29].
- Attention and eye movements share some common or overlapped brain neural circuits and control mechanisms but more importantly they can be functionally separated. Attentional shifts can be decoupled and disengaged from eye movements to freely select interesting targets without eye movements [11], [15];
- Attention plays a crucial role in the control and sequence of eye movements and visual perception. Attention filters out visual backgrounds and ensure that eye movements are programmed solely on the objects, features, locations, or groups selected by attention. Deploying attention to the targets is thought to increase search efficiency by solving the competition problem in cluttered scenes, and to ensure stable gaze maintenance and accurate and effective targeting of eye movements [15], [16].

Following these inspirations, the work presented in this paper explores a biologically-plausible framework based on visual-object based attention for integrating attentional shifts and eye movements through overlapped circuits and shared control mechanisms. The "visual object" (or visual-object in this work) concept was introduced by Pylyshyn [23] and is initially similar to the term "proto-object". In this work, we extend "visual-object" to include a proto-object (or pre-attentive object), cluster of features including locations relevant to objects, conceptual and 3D physical object, and group of objects as a formal alternative to the concept of "grouping" we introduced in [30] for integrating space and object based attention. The term "grouping" may be confused with perceptual grouping and the term "proto-object" at least lacks the power to represent physical and conceptual objects, object-based and feature-based selection, and segmenting of scenes into units with attention. Attentional shifts and eye movements (gaze shifts) are both modelled in the way of which they not only can be coupled and cooperated for flexible visual selection but also particularly can be separated or dis-coupled for their

different function roles – an important feature of human vision [11], [15]. Attention through integrated or biased competition operates upon the common underlying attentional circuits, responding with visual selection by locations, features, objects, and groups depending on the nature of attended targets and visual tasks. Attentional shifts within the attended areas (surrounding and including the fovea) to select interesting visual-objects based on visual competition and drives eye movements if necessary to explore visual-objects in the periphery of the visual field. These two kinds of shifts are integrated but separable from their functions to achieve complicated attentional selection especially in natural and disordered visual environments.

The proposed framework is novel by the following features: Gaze shifts are programmed and guided by visual-object based attention; Gaze shifts and attentional shifts are integrated but separable in functions by their biologically-plausible relationship, working coherently to achieve complicated attentional selectivity; The competition for attentional selection and eye movements is dynamic in a spatio-temporal context with local and global integration across multiple resolutions, and based on their common control circuits; The framework is inspired by recent psychophysical findings about human visual attention, eye movements and their relationship, so that it has a strong biologically-plausible theory foundation.

## 2 The Proposed Framework

### *2.1 Overview*

Sun and Fisher previously proposed and implemented a computational framework for object-based attention integrated with space-based attention through the extended concept of "grouping" [30]. Hierarchical selectivity of attention by locations, features, (structured) objects, regions, and groups of objects was demonstrated by comparison experiments with data commonly used in human psychophysical experiments of visual attention and practical performance in clustered nature scenes. Nevertheless, the groupings used in the experiments were manually segmented. The framework proposed here extends [30] to integrate eye movements with an automatical and dynamical perceptual grouping to form "visual-objects" (the replacement concept for "grouping") following each gaze shift. Attention and eye movements are constructed on common underlying attentional circuits. To achieve integrated cooperation and also functional distinguish between attentional shifts and gaze shifts, the mechanisms of retinal-like sensor, dynamically spatio-temporal visual object-based saliency mapping, temporary Inhibition Of Return (tIOR) (with short-term memory) and attention-driven switch are incorporated, as illustrated in Figure 1. Gaze

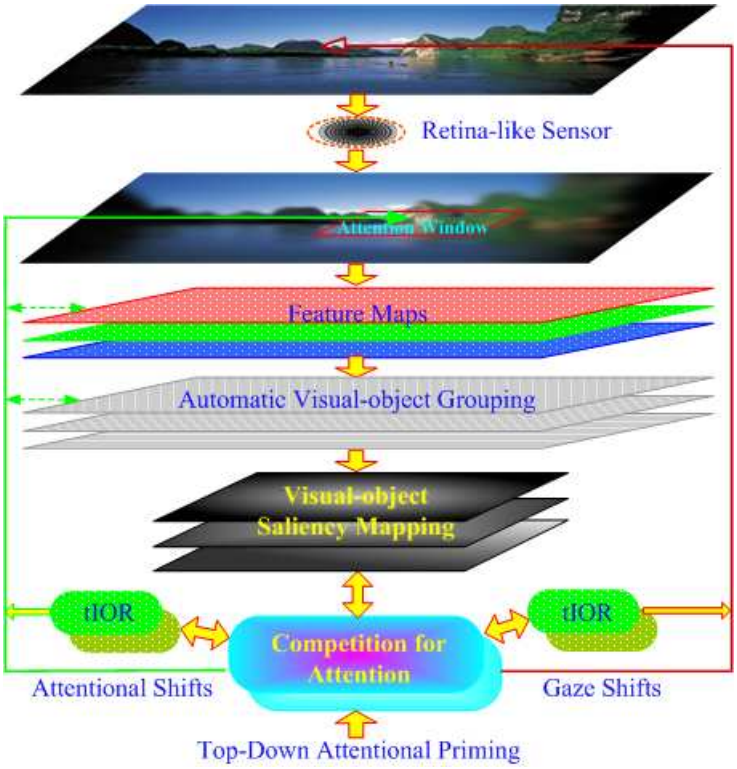


Fig. 1. Overview illustration of the proposed framework.

shifts cause the retinal-like sensor to create a series of foveated images due to eye movements over time. Automatic perceptual grouping is then dynamically formed in each foveated image. Spatio-temporal visual-object based saliency mappings are built correspondingly. Through Winner-Take-All (WTA) and tIOR, a winning visual-object captures an eye movement to shift the fovea into its most salient area and following this, attention shifts within the high acuity area (attention window) to select interesting visual objects. The next target position of an eye movement will be produced through the competition for attention between unattended visual objects outside the current attention window. The engagement and disengagement of attention from eye movements is controlled by the mechanism of attention-driven switch (2. The main extensions in this framework are described in the following.

### *2.2 Retina-Like Sensor and Foveated Image*

Human vision makes frequent use of discrete saccadic fixations to explore the visual world and to help visual attention quickly gather interesting information critical to current visual behaviour. Space variant sensing mimics the capability of the human retina where the fovea with finer spatial resolution is used to observe interesting objects in more detail and the periphery with increasingly coarser resolution is used to rapidly detect potentially interesting

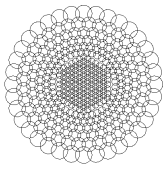

Fig. 2. Structure of the retinal mask used to derive a log-polar image

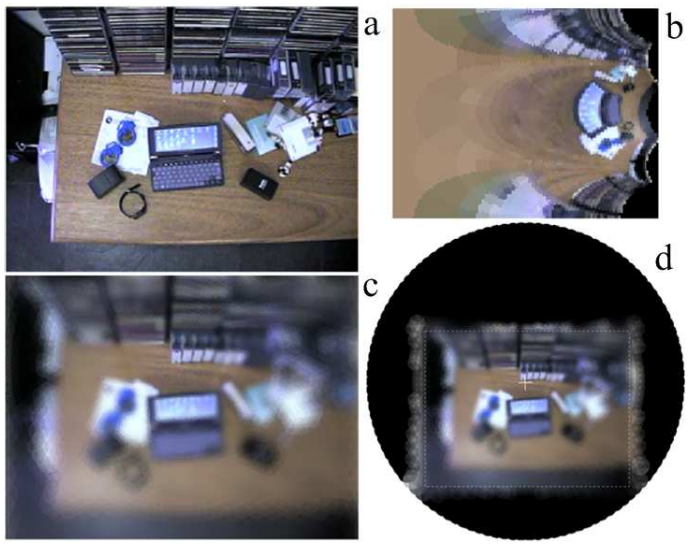


Fig. 3. An example of how to create a (reconstructed) foveated image. a: input image; b: log-polar image (magnified); d: diagram that shows the foveation center (shown by a cross), the retinal mask area (within the large circle) and the clipping rectangle (dashed white rectangle) used to reconstruct the foveated image c.

objects in the field of view. The retina-cortical mapping in the human visual system can be simulated through a log-polar mapping which has been widely used in machine vision.

The retina-sensor here first uses a mask (Figure 2) to sample the input image to produce log-polar images and then generates foveated images in the Cartesian space through a re-mapping procedure (see [7] for the details). Figure 3 shows an example of this pair of processes. This sensor simulates human retinal imaging and has properties: uniform fovea and log-polar periphery, overlapping and Gaussian weighting over receptive fields and hexagonal neighborhoods. This approach can reduce the implementation complexity and is faster than directly extracting features in the log-polar images. Using log-polar images for space-variant feature extraction has some well-known advantages but also has limitations, e.g., designing new image processing operators, complicating object-based perceptual grouping because the size and shape of image features change radically as the gaze shifts. The possible approaches to overcome the limitations caused by log-polar transforms are to make use of the Mellin-Fourier transform or connectivity graph [2], learning-based neural networks [7] and the inverse mapped log-polar image with pyramid algorithms [2] as

adopted in this work.

### *2.3 Attentional Window*

Many findings [17, p. 27-39] suggested that there exists a relatively sharp boundary between an attended area and its surround and the attended area is movable. The "zoom-lens" metaphor furthermore suggests an attended area of variable size and shape with high clarity at the centre and gradually decreased clarity from the centre to their periphery. However, the research on a generally accepted account for the attended area (also called attention window) is still open. Inspired by the above suggestions and for simplicity, a square attention window (with size $256 \times 256$ pixels) is adopted in this work to show how (covert) attentional shifts and (overt) gaze shifts work together in this proposed visual object-based selection framework. The attention window is assumed to centre on the eye fixation position which can be either the centre of mass or the most salient location of an attended/fixated visual object.

### *2.4 Transition Between Attentional shifts and Gaze Shifts*

The models shown in Figures 1 and 4 implement two kinds of shifts: 1) attentional shifts within the attentional window, a high acuity area enveloping the fovea, to perform visual selection of interesting visual objects; and 2) attention-driven gaze shifts (e.g., saccadic eye movements) outside the window to assist attentional selectivity in the whole field of view. At any time, the shift made by attention or by gaze depends on the current state of visual-object competition for attention, visual-object position, and visual tasks. If a visual object that lies outside the attentional window wins the competition, a gaze shift will be triggered by attention. Otherwise, when a visual object within the attentional window wins the competition, attention will be disengaged from eye movements and shifts to select that visual object. The transition between these two kinds of shifts allows a visual system flexibility to deal with complex visual selection tasks.

### *2.5 Primary Feature Extraction and Feature Maps*

From each foveated image obtained with a saccadic eye movement, the color, intensity and orientation features and corresponding feature maps are created using the same approach as in [30] that employs multiple pyramids to build a pyramidal saliency mapping to achieve coarse to fine attentional selection without using a space variant sensor. Here we do not need to use pyramidal

feature maps but only require the finest resolution maps, as the retinal sensor can provide a more natural and biologically-plausible way. It is also noted that these feature maps are obtained dynamically from foveated images with gaze (retina-sensor) shifts.

### *2.6 Automatic Segmentation and Perceptual Grouping*

Visual objects (or groupings) are dynamically formed as a hierarchy of grouped regions, objects and other components from each foveated image created with each saccadic eye movement, because scene context may change due to gaze shifts. The automatical grouping approach adopted in this work consists of two steps. After a gaze shift from its previous foveal location to a new foveal location, a new created foveated image is automatically segmented into regions by using EDISON (Edge Detection and Image SegmentatiON) approach proposed by B. Georgescu and C. M. Christoudias [25]. This segmented image is then processed by a graph-based partitioning/grouping method inspired by Y. Haxhimusa and W. G. Kropatsch's research [8], to construct hierarchical groupings (or visual objects). We found, by adjusting the parameters, the combination of these two methods can generate acceptable and grouped structures in many scenes, though the results are not very ideal compared with using a manual grouping approach. The research here, however, is not focused on the studies of ideal hierarchically perceptual grouping approaches. Rather, these generated groups are just used to more clearly show how eye movements and attentional shifts work together to perform human-like attention tasks. Figure 9 and 14 showed the groupings resulted from the scenes (Figure 5) used in this paper.

### *2.7 Visual-Object-Based Saliency Mapping*

We first calculate the saliency of any visual-object formed by the automatically segmented foveated image at time $t$, and then map it into the corresponding visual-object-based saliency mapping. Suppose $\Re$ is any visual-object and $S(\Re)$ is its saliency obtained by using the same formulae in [30] for calculating a grouping's saliency. $SaliencyMap(t)$ is a visual-object-based saliency mapping at time $t$, and $S_{i,j}(\Re)$ is the saliency of visual-object $\Re$ mapped into $SaliencyMap(t)$ at the position $(i, j)$. $SaliencyMap(t)$ can be represented as:

$$SaliencyMap(t) = \begin{bmatrix} \cdots & \vdots & \cdots \\ \cdots S_{i,j}(\Re) & \cdots & \\ \cdots & \vdots & \cdots \end{bmatrix}_t \tag{1}$$

The saliency of a visual-object is the integrated saliency of all its hierarchically structured components and therefore represents how a visual-object stands out from other visual-objects at the same hierarchical lever by its integrated saliency.

When a scene is sampled by a foveal sensor over time, every location in the scene is observed at multiple resolutions derived from the combination of nonuniform sensing during a series of gaze shifts. Different and partially overlapping foveated images are therefore produced from the fixation shifts, resulting in saliency of each location varying in this spatio-temporal context (see Figure 7 for an example). The visual system must deal with the integration of multiple saliency mappings in a space-time context. Moreover, each foveated image contains multiple resolutions and a particular location in the scene is observed at different resolutions due to multiple gaze shifts and accordingly has different saliency when viewed from the different fixation positions over time. It is reasonable to integrate all of the previous saliency mappings over time when building the saliency mapping at the current time. Because at each time a specific location can only belong to a single resolution, combining saliency at this location over time is actually equal to combining saliency from multiple resolutions from gaze shifts. In this way, we have the following approach to build a spatio-temporal saliency mapping at a given time.

Suppose $RawSaliencyMap_{\Phi}(t)$ is the visual-object-based saliency mapping created at time $t$ (see Eq. 1), $\Phi = global$ and $\Phi = local$ denote the saliency mapping outside and within the attentional window respectively. $SaliencyMap_{\Phi}(t-1)$ is the fused saliency mapping obtained from the last foveated image and after any attention shifts at time $t-1$. Then the new fused visual-object-based saliency mapping at the current time $t$ is built as:

$$
\begin{aligned}
&SaliencyMap_{\Phi}(t) = \alpha SaliencyMap_{\Phi}(t-1) + (1-\alpha) RawSaliencyMap_{\Phi}(t) \\
&= \begin{cases} \alpha SaliencyMap_{local}(t-1) + (1-\alpha) RawSaliencyMap_{local}(t) & \Re \in AW \\ \alpha SaliencyMap_{global}(t-1) + (1-\alpha) RawSaliencyMap_{global}(t) & \Re \notin AW \end{cases}
\end{aligned} \tag{2}
$$

where $SaliencyMap_{\Phi}(0) = RawSaliencyMap_{\Phi}(0)$, $\alpha$ is a constant $\in (0,1)$, $AW$ indicates the attention window, $\Re$ is any visual-object. At any time $t$, $SaliencyMap_{local}(t)$ and $SaliencyMap_{global}(t)$ are contained in the same saliency mapping $SaliencyMap_{\Phi}(t)$. $SaliencyMap_{global}(0)$ and $RawSaliencyMap_{global}(0)$ initially equal at time $t = 0$ are used to generate the first gaze shift from the initial fixation place to a new place at time $t = 1$. Attention shifts occur after time $t = 1$. The parts of the mapping within (indicated by $\phi = local$) and outside (indicated by $\phi = global$) the attentional window are created at the same time. Within the window a small Gaussian weighted distance (e.g., $\sigma \leq 5\%$) is used for visual-objects' saliency computation [30] which guarantees

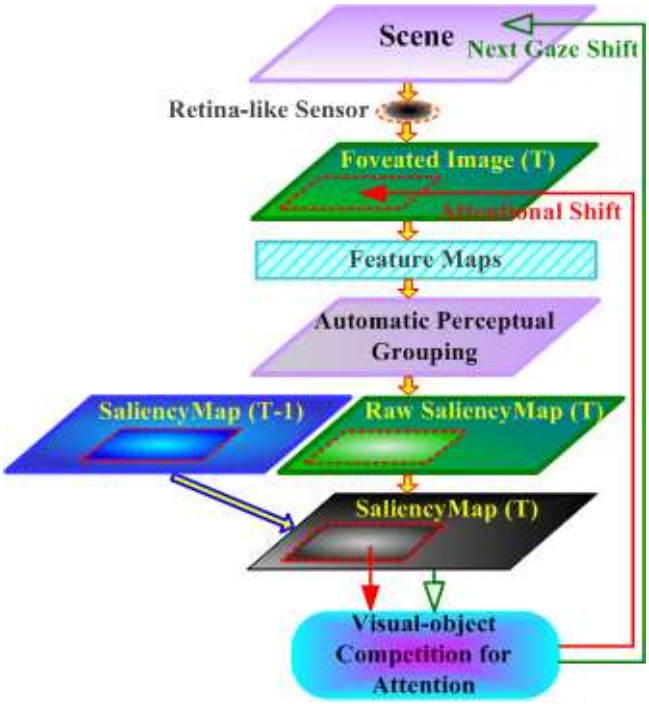


Fig. 4. The routes of how to build the spatio-temporal visual-object saliency mapping and how two kinds of attentional shifts and gaze shifts occur within and outside the attentional window (red dashed frame). Red solid arrow indicates an attentional shift and green open arrow indicates a gaze shift.

that the competition for attention selection is confined to a local area. Outside the window a larger Gaussian weighted distance (e.g., $\sigma \geq 20\%$) is used for visual-objects' saliency computation that guarantees that the competition for a saccade covers the whole field of view.

The above saliency mapping idea can be further illustrated in Figure 4. The calculation shown in Eq. 2 provides the temporal integration of the raw saliency across multiple fixations and attentional shifts over time. After this temporal integration, the competition in the scene across multiple resolutions can be reasonably reflected in the new visual-object saliency mapping.

In Figure 4, $SaliencyMap(t-1)$ denotes the combined or fused visual-object-based saliency mapping at time $t-1$. $RawSaliencyMap(t)$ is the saliency mapping produced directly from the new foveated image after a gaze shift in the scene at time $t$. $SaliencyMap(t)$ is the new (combined) saliency mapping created at time $t$ by integrating $SaliencyMap(t-1)$ and $RawSaliencyMap(t)$.

Based on this updated saliency mapping, at time $t$ attention shifts (shown by the red solid arrow in Figure 4) within the attentional window to select the winners of visual-objects which compete for attention. Inhibition (by Eq. 3) is then applied to the previously attended visual-object to prevent attention from immediately returning to this visual-object. After attention shifts, the saliency of the suppressed visual-objects starts to increase again in the saliency

mapping $SaliencyMap_{\phi}(t)$. Visual-objects outside the window, and the parts of visual-objects across the window boundary compete for the next saccade.

### *2.8 Temporary Inhibition of Return (tIOR)*

Inhibition of return (IOR) [22] is a transient bias mechanism which prevents attention from instantly returning to a previously attended target in a short time period. It involves temporal aspects of visual selection. A visual system requires sufficient dwell time to accomplish a visual selection. On the other hand, after a minimum avoidance time, previously selected objects should be allowed to regain visual selection. This is especially useful for a vision system when exploring complex scenes that normally contain hierarchically structured objects that need to be reattended for some further processing. Some findings have shown that there is a close link between IOR and saccade programming [9]. Important evidence also shows that IOR is partly object-based and moves with objects to the new positions [33]. It was reported that IOR can operate simultaneously over several objects, that is, multiple previously selected loci/objects can be inhibited at once [35]. Recent studies have shown that even simple visual tasks elicit many saccades that often repeatedly visit the same objects, and visual attention required by saccades is guided by a short-term memory system that facilitates the rapid refixation of gaze to recently foveated targets [19].

IOR has been broadly used in many computational models of attention or eye movements but most of them only used a simple suppression version of IOR and have not considered the above complex properties of IOR. IOR's spatio-temporal dynamics that enable attention to reattend an object and multiple object-based properties, however, are important and necessary for a vision system to deal with complicated visual tasks effectively. The IOR mechanism is proposed here in a temporal context and termed "temporary Inhibition Of Return" (tIOR), which is used to temporarily prevent both attention and gaze from immediately returning to the last accessed visual-object. After attention selects a visual object in the attentional window and is going to shift, the attended object must be transiently inhibited so as to avoid regaining attention immediately. As the attention window moves with a saccade, this inhibition is correspondingly applied to each attended visual-object. Because a fixated visual-object is located within the window and is attended first, it is also suppressed when a saccade shifts to it. The inhibition of return is thus applied to both attentional shifts and saccadic shifts.

Let $\Re$ be an arbitrary and unattended visual-object or sub-visual-object belonging to a structured or grouped visual-object within the attentional window at time $t$, and $S(\Re)$ be its saliency value $\in SaliencyMap_{local}(t)$. We can use

$t = 0$; while (the given goal is not reached)
{ create $RawSaliencyMap_{\Phi}(t)$ and recreate $SaliencyMap_{\Phi}(t)$ through Eq. 2
and "tIOR rules" for the saliency mapping update;
(Note: $SaliencyMap_{\Phi}(t) \supset SaliencyMap_{global}(t)$ and $SaliencyMap_{local}(t)$);
$S_X = Max(SaliencyMap_{global}(t))$ ($X$ is the winning unattended visual-object);
saccade to $X$ and create an attentional window at the fixation position;
$t = t + 1$; $t' = t$; $i = 1$; $end = FALSE$;
while($i <= n$ and not $end$) ($n$: total visual-objects in an attentional window)
{attention selects an unattended visual-object $i$ based on $SaliencyMap_{local}(t')$;
suppress visual-object $i$ within $SaliencyMap_{local}(t')$ using Eq. 3;
$t' = t' + 1$;
if $S_{in}(n - i) < \varphi \cdot S_Y = Max(SaliencyMap_{global}(t))$
($\varphi > 0$ is a constant, $Y$ is unattended and outside the AW)
{ saccade to visual-object $Y$; $end = TRUE$}
else $i = i + 1$; }
$SaliencyMap_{\Phi}(t) = SaliencyMap_{\Phi}(t')$; }

**Algorithm 1.** The algorithmic description of *temporary Inhibition Of Return*

the following suppression function $SUPP(SaliencyMap_{\Phi}(t), \Re)$ to suppress $\Re$ if it has been attended:

$$SUPP(SaliencyMap_{\Phi}(t), \Re) = \begin{cases} S(\Re) = \Theta \cdot (1 - \exp(-d^2/D^2)) & \text{if } \Re \text{ has already been attended;} \\ S(\Re) & \text{otherwise;} \end{cases} \quad (3)$$

where $\Theta \geq 0$ is an inhibition threshold and can be a real constant close to 0 or simply set to be 0. $D$ is a scale measure of visual-object $\Re$ and can simply take the count of members of $\Re$ as a measure. $d$ is the distance from the center of mass of visual-object $\Re$ to the gaze fixation location at time $t$. The above suppressing function can also be made by a simple way that uniformly decreases the entire saliency activities of an attended visual-object below a given threshold (e.g., 0) at one time. To complete tIOR, we have the following "tIOR rules" to update the saliency mapping:

1) If some parts of a visual-object $\Re$ have been attended and suppressed using Eq. 3 but some other parts have not been attended, the saliency of visual-object $\Re$ will be updated using Eq. 2;
2) Otherwise, visual-object $\Re$ will not take part in the saliency updating.

Given the spatio-temporal visual-object-based saliency mapping within an attention window, attention covertly selects the salient visual-objects within the window over time. When will attention shift to a salient visual-object located at the periphery? An ideal solution to this problem may involve complicated top-down and bottom-up reasoning. We use here a simple way to achieve this switch from attentional shift to a saccade. We assume that within an atten-

tion window there are $n$ hierarchical visual-objects that compete for attention. When the gaze is maintained on its position, attention is disengaged from this eye movement and shifts to select interesting visual-objects over time. After attention shifts to the *ith* visual-object or part of a visual-object, if the most salient visual-object among visual-objects outside the attentional window is unattended and more salient than the weighted sum of saliency of all unattended $(n - i)$ visual-objects within the window, this visual-object wins the competition and gains attention. Attention then drives a gaze shift and is engaged with this eye movement bringing the fovea on that visual-object. This simple approach can effectively avoid exhaustive search within an attended area and is easily be implemented.

Suppose $S_{in}(n-i)$ is the sum of saliency of remaining $n-i$ unattended visual-objects within an attention window which includes $n$ visual-objects in total at time $t$. The other $i \geq 0$ visual-objects have been attended and suppressed. Let $Y$ be a visual-object outside the attention window. The algorithm implemented for temporary inhibition of return (tIOR) is shown in Algorithm 1.

### *2.9 Switch Between Attentional Shifts and Gaze Shifts*

Switching between attentional shifts and gaze shifts involves engaging and disengaging of attention from eye movements. The next saccadic target is selected from the following "switch rules":

1) an unattended visual-object, or a partially attended visual-object that is the most salient of visual-objects crossing the boundary of the attention window and is not less salient than the most salient visual-object outside the window; or
2) the most salient and unattended visual-object following Algorithm 1 outside the attention window if the above is not available.

The switch between attentional shifts and gaze shifts is controlled by the mechanism of attention-driven switch given in Algorithm 2. When the fovea is fixated at a position, attention will disengage (or discouple) from this eye movement and shift freely to select interesting visual-objects around the fovea. When attention needs to attend to a new visual-object at the periphery or the unattended remainder of an attended visual-object that lies across the attentional window, attention will drive a gaze shift and engage with it bringing the fovea to the new fixation position, so that attention can perform fine selection at higher resolution. Attentional shifts and gaze shifts are therefore modelled upon the shared underlying circuits but at different function levels to fulfill their own roles while working together to accomplish complex visual selectivity.

1. Assume the fovea initially fixates on a random position (e.g., the centre) of the input scene to create a foveated image by the retina-like sensor;
2. Construct hierarchically structured visual-objects from this foveated image;
3. Create the visual-object-based saliency mapping;
4. Attention is engaged with a saccade and drives it to shift the fovea into the winning visual-object;
5. Produce a new foveated image at the new fixation position;
6. Construct the new hierarchically structured visual-objects;
7. Create a new visual-object-based saliency mapping (Eq. 2 and "tIOR rules");
8. The gaze is maintained on its position and the competition for visual-object-based attentional selection is triggered;
9. Attention is disengaged from the gaze and shifts within the attentional window to select unattended visual-objects (using tIOR (Eq. 3) to suppress) until an unattended visual-object (following "switch rules") wins the competition;
10. Update the visual-object-based saliency mapping after each attentional shift;
11. A new saccade is triggered and brings the fovea to a new position according to "switch rules";
12. If (all unattended visual-objects in the input scene are attended or a given goal is reached), Go to step 13. Otherwise go to step 5;
13. Stop.

**Algorithm 2.** Mechanism for switch between attentional shifts and gaze shifts


Fig. 5. Natural scenes

## 3 Results

### *3.1 Behaviour and Performance in Natural scenes*

We firstly present two natural scenes (shown in Figure 5) to show how attentional shifts and saccadic eye movements work together for efficient and co-

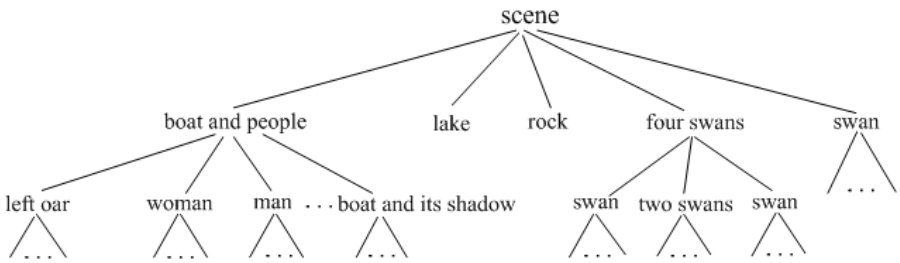


Fig. 6. A possible ideal visual-object hierarchy in scene 1

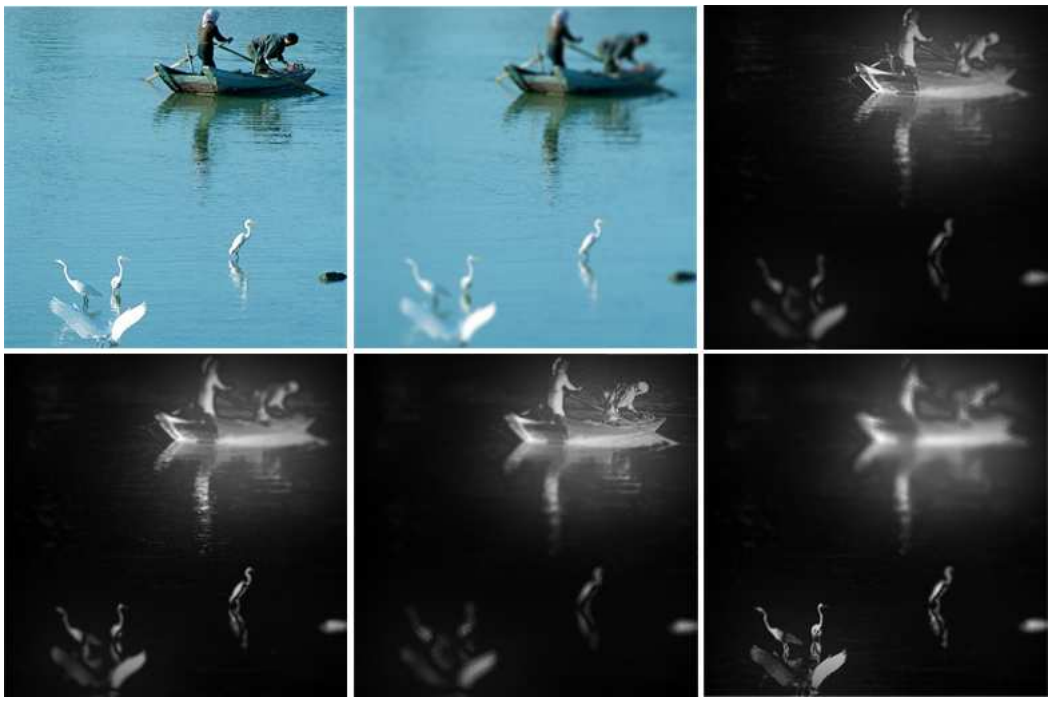

Fig. 7. Top left: a natural scene 1; Top middle: initial foveated imaging. The rest are "location-based" saliency maps computed from the first gaze shift to the one before last shift, used for the comparison with the visual-object-based saliency mapping (Figures 10 and 11). Note saliency (brightness) of locations varying following gaze shifts.

herent visual-object based selection. The framework performed 7 saccades in scene 1 (Figure 8) and 6 saccades in scene 2. With each saccade, the framework performed several attentional shifts to achieve hierarchical selectivity (i.e., multi-level selection) by location, object, region or group of visual-objects. To explain a hierarchy of visual-objects, we use a graph in Figure 6 to illustrate a possibly ideal partition in scene 1. In this supposed ideal segmentation, there would be four top level visual-objects which are hierarchically structured: the region of lake, the hierarchically structured object of boat and two people, the group of swans and the single object of rock. The visual-object of boat and people also includes several structured and overlapped sub-visual-objects: a woman, a man, the boat itself, oars and their shadows reflected in the water etc. A similar hierarchical structure exists for the people and group of swans. Figure 7 shows several "pixel" or "location" based saliency maps for the pur-

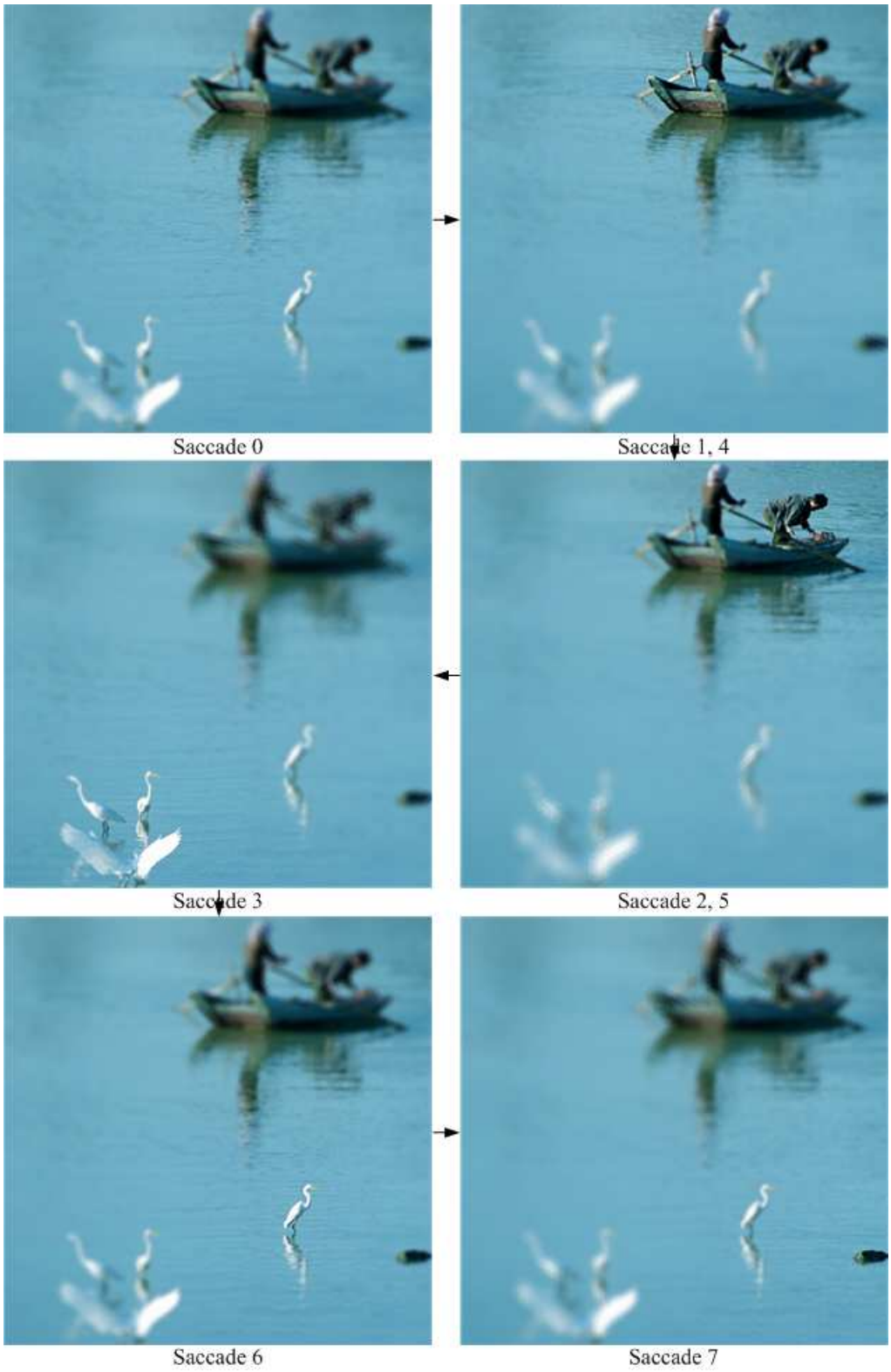


Fig. 8. Retinal imagings derived from saccadic eye movements in scene 1. Note saccade 4 and 5 for re-attending processes due to tIOR.

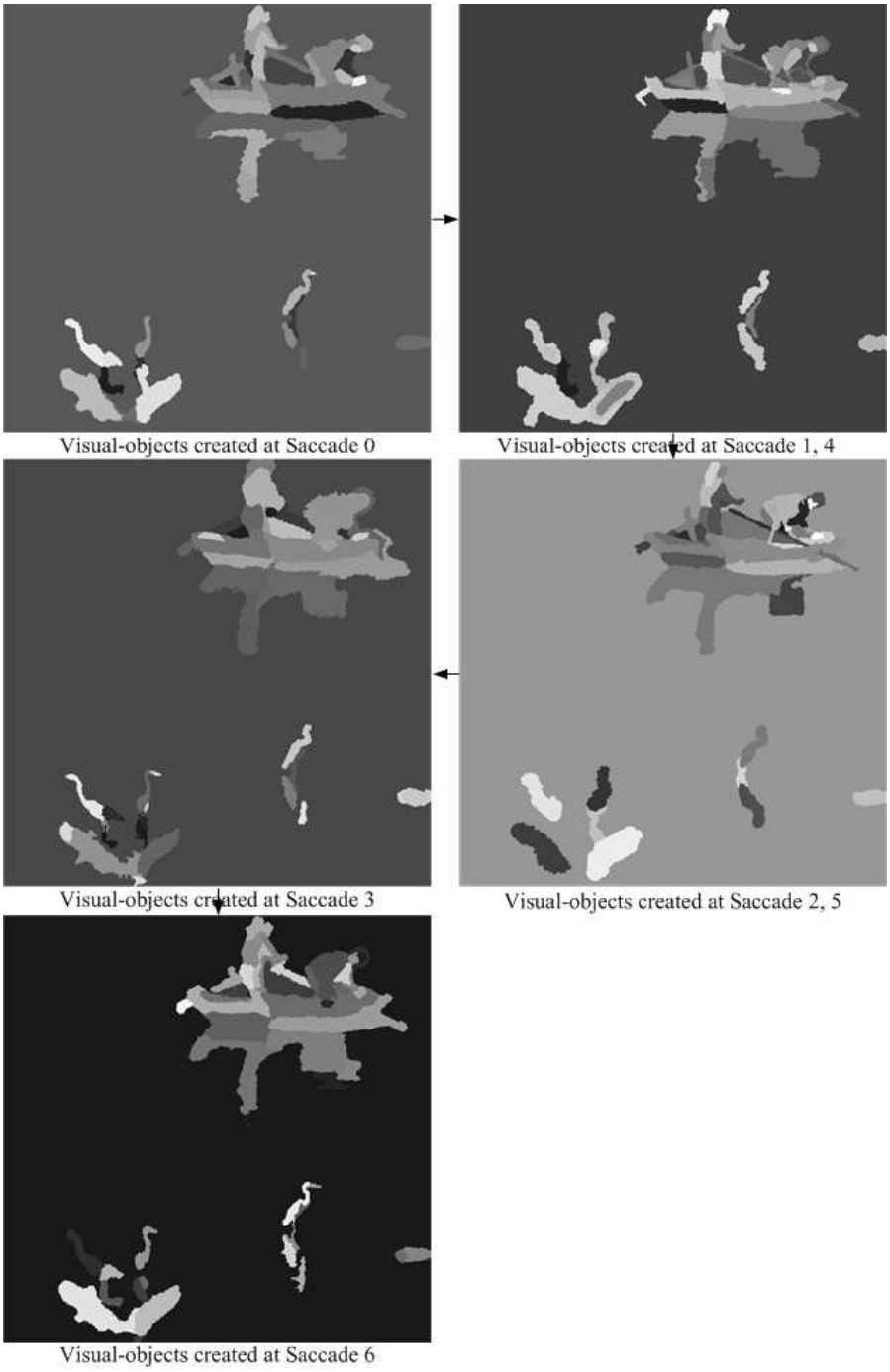


Fig. 9. Hierarchical visual-object structures dynamically change over time due to gaze shifts in scene 1.

pose of comparing with visual-object-based saliency maps (Figures 10 and 11) used in this work. However, the work presented in this paper actually used automatical segmentation and perceptual grouping (see Section 2.6) to obtain dynamic visual-objects based on each foveated image created after each gaze shift. Following this, the visual-object-based saliency mapping also dynamically changes. Figure 8 shows foveated images created from each saccadic shift in scene 1 and correspondingly segmented and formed dynamic visual-objects

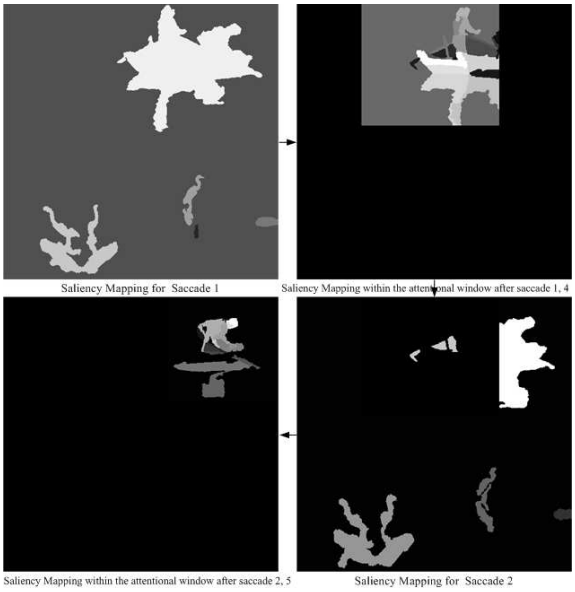


Fig. 10. Spatio-temporal saliency maps obtained from the natural scene 1 during eye movements and attentional shifts. Different saliency strengths of visual-objects are shown in different grey scales where the brighter is more salient. Note the change of saliency mapping caused by tIOR.

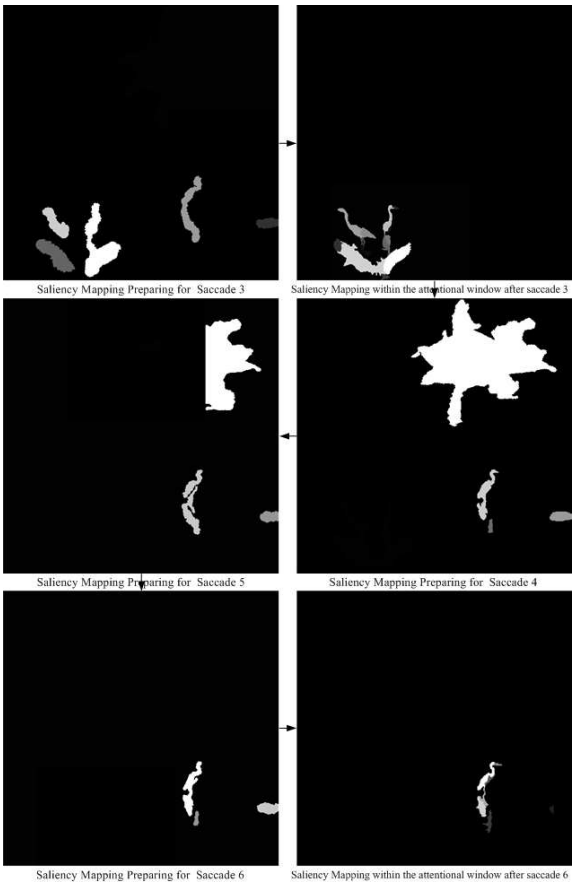


Fig. 11. Spatio-temporal saliency maps continued from Figure 10.

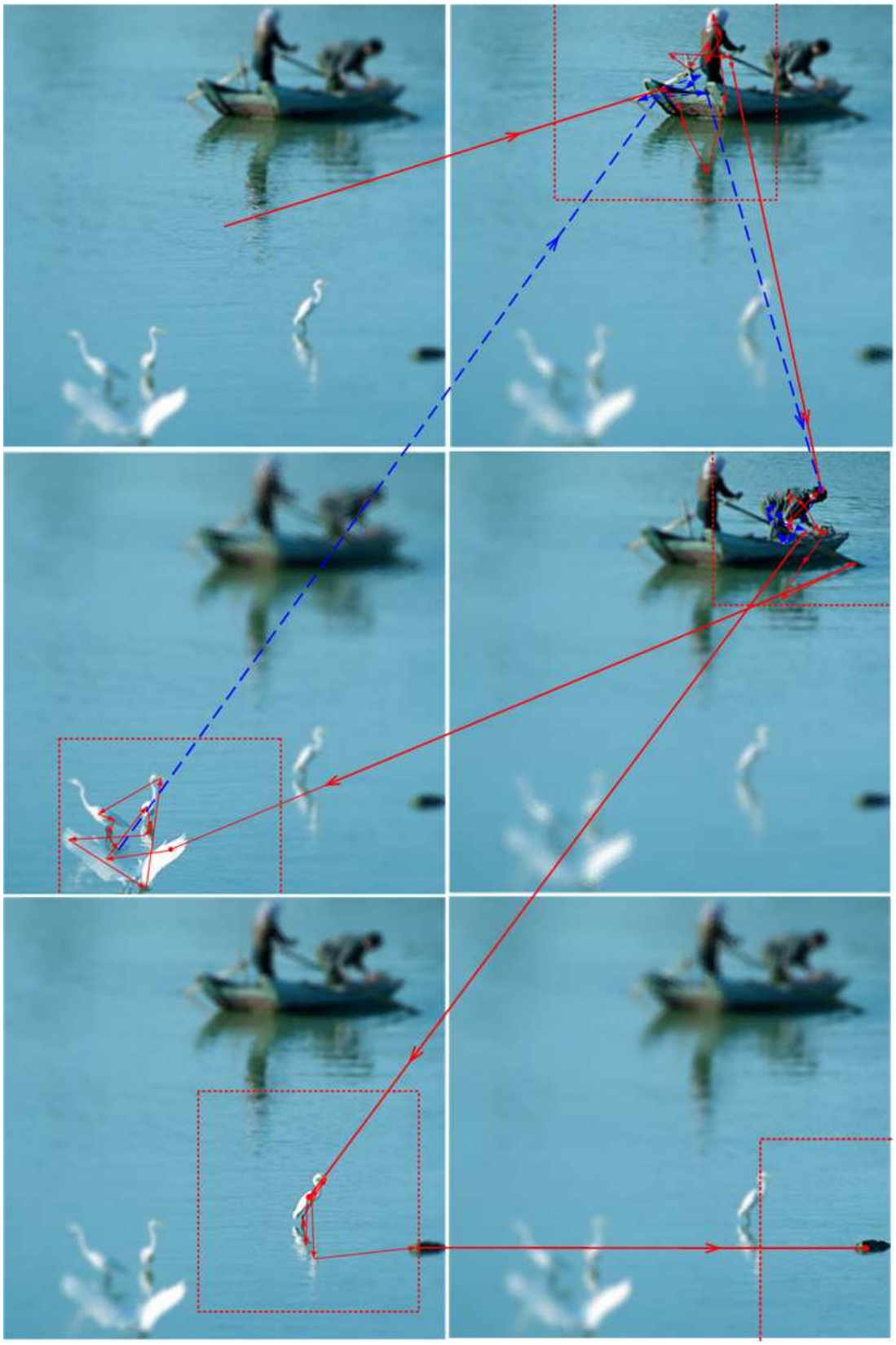

Fig. 12. Both attention-driven eye movements and attentional shifts are shown. Red broken lines: attentional windows; Arrows within the windows: (covert) attentional shifts. Dashed blue arrows indicate attentional shifts to previously unattended visual-objects due to tIOR; Arrows outside the windows: gaze shifts. Dashed blue arrows indicate gaze shifts due to tIOR.

are shown in Figure 9. Similarly, Figure 15 and Figure 14 show the foveated images and relevant formed visual-objects from scene 2.

The first foveated image generated from scene 1 is shown at the top left image in Figure 8. Before a saccade is ready to launch, this foveated image is partitioned by the automatic perceptual grouping process to form visual-objects (top left image in Figure 9) and the visual-object-based saliency mapping (upper left image in Figure 10) is then produced. As visual-objects compete for attention, attention is engaged with a saccade and drives the gaze shifting to the winner, i.e., the most salient visual-object here. The fovea is therefore brought to a new location and a new foveated image (top right image in Figure 8) is produced following this saccadic eye movement. Correspondingly, visual-objects are re-formed (shown in Figure 9) and their relevant visual-object-based saliency mapping (The top row images in Figure 10 show this dynamic updating.) is updated over time according to this fovea shift. When the fovea is maintained on its position, the dynamically reformed visual-objects within the attentional window compete for attentional selection. Attention is disengaged from this eye movement and shifts to select the winning visual-objects. The scanpath of saccadic shifts and attentional shifts are shown in Figure 12. After several attentional shifts monitored by the tIOR mechanism, a visual-object outside the current attentional window may win the competition for attention. A new saccadic eye movement is then programmed and ready to shift. The previously attended and suppressed visual-objects may take part in later competitions for attention and may possibly gain re-attending if their saliency rises to a significant level from previous suppression level (e.g., the fourth saccade and sixth saccade which are shown in Figure 12 by the blue arrows). This results in some visual-objects, e.g., the boat with people, being re-attended twice. The dynamic update of visual-object saliency mappings over time during gaze shifts and attentional shifts are given in Figures 10 and 11.

It is clear that visual-objects and their saliency dynamically vary over time when the fovea position shifts. The competitive capacity of a visual-object to gain attention varies with this dynamic changing. Attentional shifts and gaze shifts are clearly shown in their differently functional pathways though they are built on the shared and overlapped control network circuits. This kind of biologically-compatible behaviour about the relationship between attention and eye movements is achieved naturally.

### *3.2 Relation to Other Works*

Until now, we have not found other computer vision research similar to ours, which clearly modelled attentional shifts for visual-object-based selection and

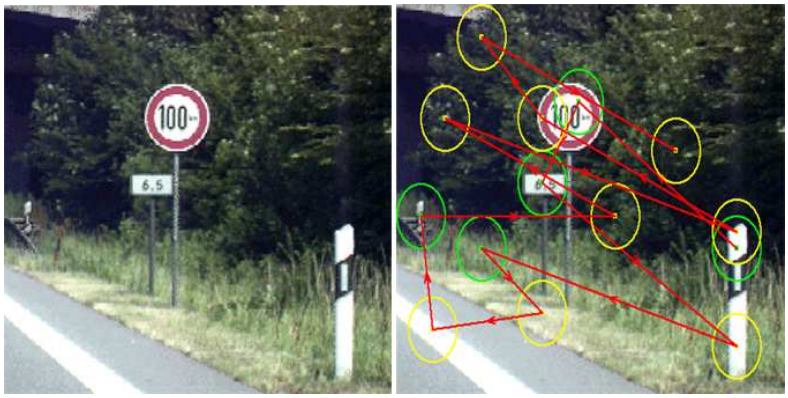

Fig. 13. Left: scene 2 (originated from [13]); Right: attentional shifts obtained from [13].

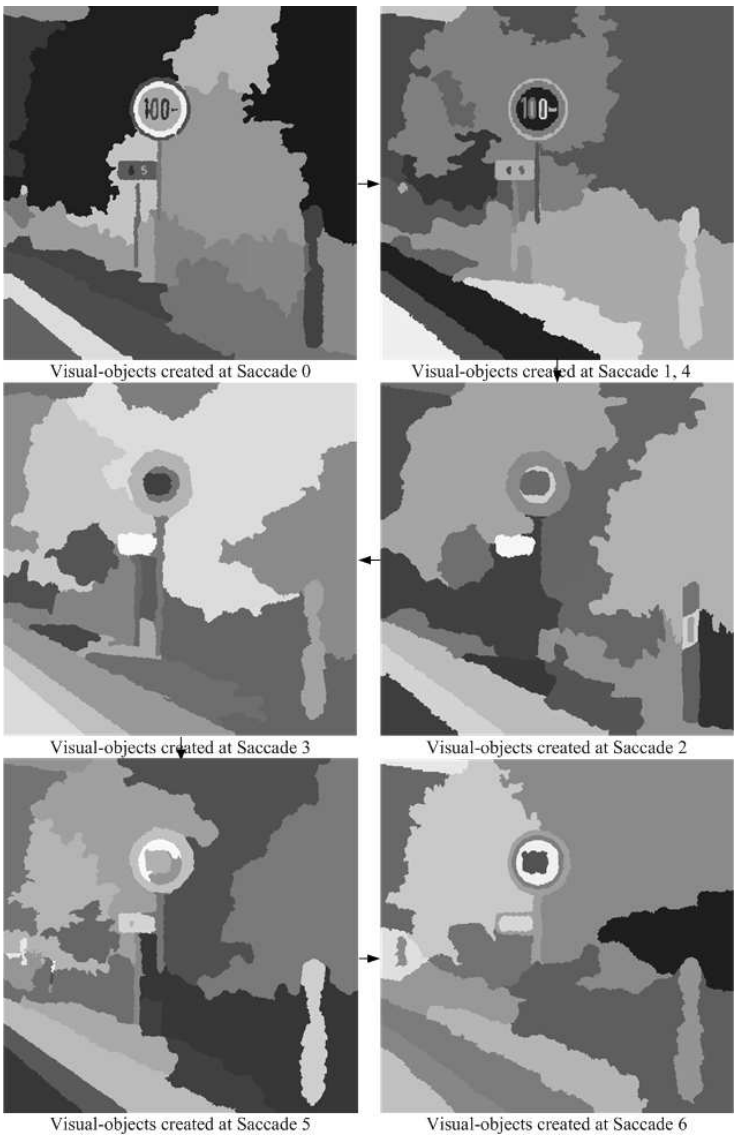


Fig. 14. Visual-objects automatically formed from different fixation position over time due to gaze shifts.

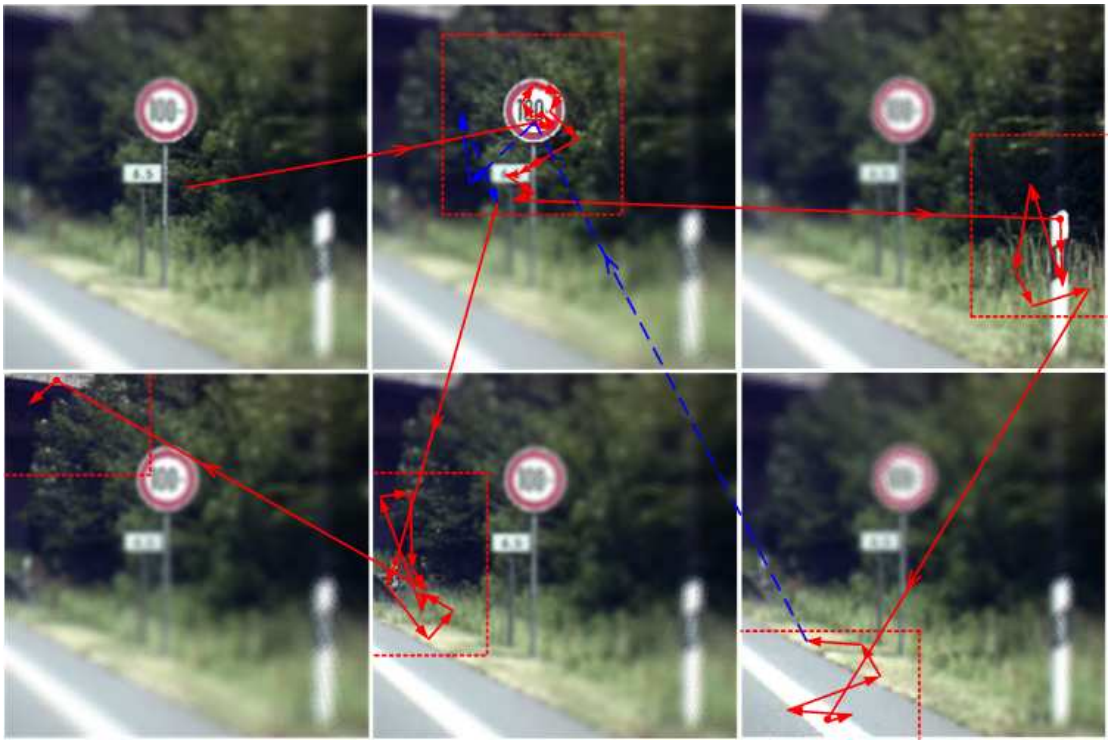

Fig. 15. The scanpath of gaze shifts and attentional shifts.

eye movements as a support role for complicated attentional selection within a coherent system in a biologically-plausible way. The biological-plausibility of our approach to integrate multiple attentional selection by locations, objects and groups through common attentional circuits has already been demonstrated by comparing its behaviour with psychophysical data resulted from research on human visual attention [30]. It fixes many problems of previous attention models and practical performance has been demonstrated by comparing its performance with other computer vision works in many natural scenes [31]. The work presented here, therefore, will not repeat these similar demonstrations and usefulness for modelling attention. Rather it is focused on demonstrating how attentional shifts and eye movements can be modelled in a computational and biologically-plausible way to work coherently to deal with complicated selection in clustered natural scenes.

The natural scenes and corresponding results in Figures 13 and 16 adopted from [13] show how attentional shifts and gaze shifts work together to efficiently select more interesting visual-objects with fewer shifts, fewer non-meaningful locations or regions irrelevant to visual-objects and to achieve effective visual search by attentional shifts between fovea-periphery and periphery-periphery. The right image in Figure 13 shows previous results from [13], it can be seen that some obvious objects, e.g., numbers "100" and "6.5", the road and etc. in scene 2 were not attended, while the trees were redundantly selected at least 4 times. These problems are overcome in our approach (shown

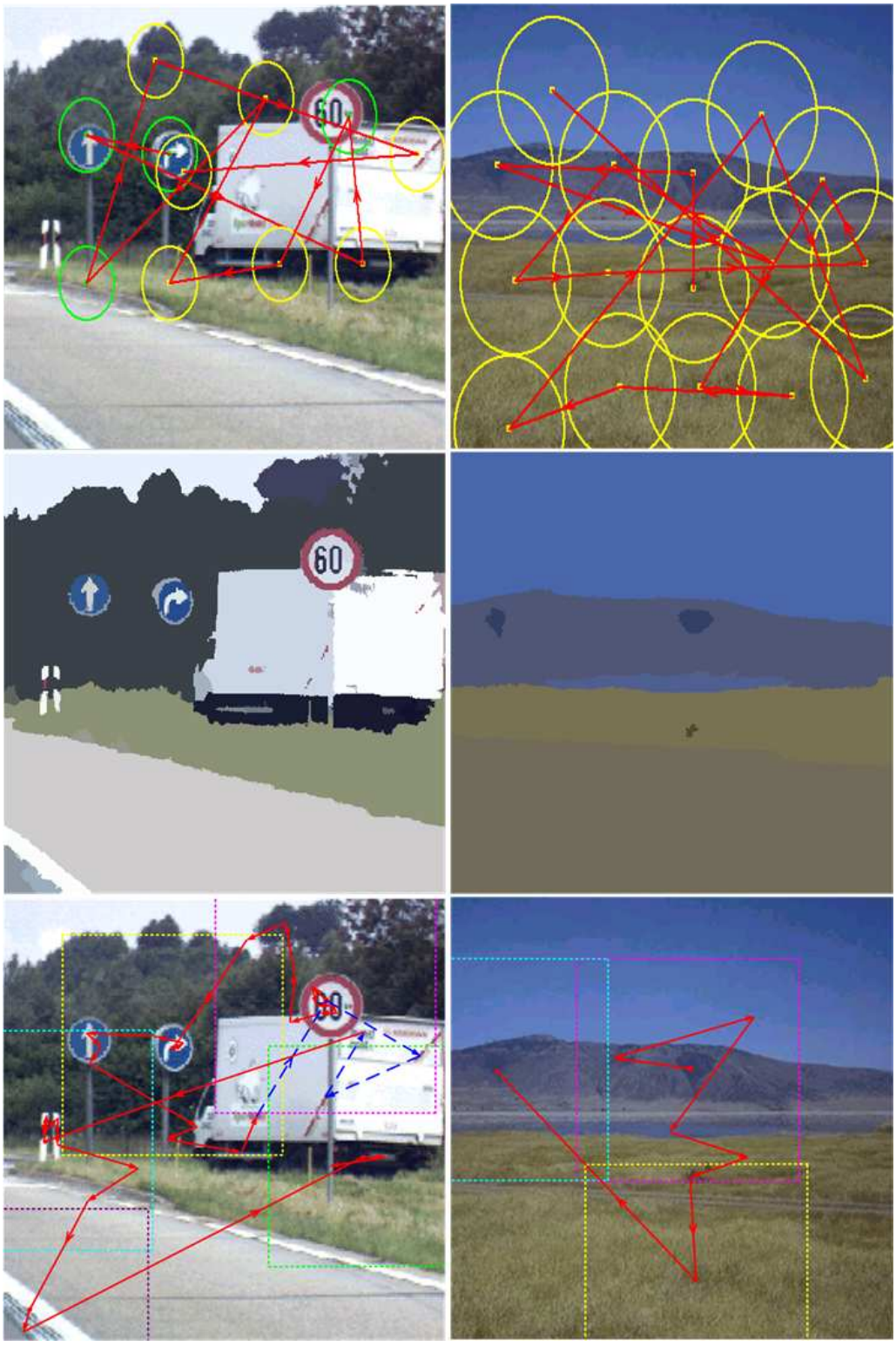


Fig. 16. The first row shows the results obtained from [13] and the third row shows the corresponding scanpaths of gaze shifts and attentional shifts resulted from our work. The second row shows visual-objects constructed from the first foveated images. Arrows here are similar to their implications in Figure 12, while we collect all attentional shifts and gaze shifts in different foveated images shown in one original image for the purpose of illustration. Coloured attentional windows in the figures indicate different saccades.

in Figure 15). Figure 14 shows the dynamically formed visual-objects automatically segmented during eye movements. In Figure 16, the first row shows previous results from [13] and the third row shows our corresponding integrated scanpaths of attentional and gaze shifts. In the top left scene, some salient objects, e.g., two white pillars and number "60" which were not selected are corrected in our work. The top right scene shows that 16 locations in this simple scene were selected, but actually 8 attentional shifts are enough to select all salient visual-objects and others are redundant and nonsense. In addition, in the large scale scene, attention needs to employ eye movements to attend interesting objects in the periphery of the visual field because the fovea is limited in range and can only fixate limited objects at one time. This kind of switch between attentional shifts for visual selection and gaze shifts for supporting attention has been implemented only in our current work.

## 4 Discussion and Conclusion

This paper proposed a novel framework based on integrated competition of attention to integrate attentional shifts and eye movements in a biologically-plausible approach. Attentional shifts for multiple visual selectivity and gaze shifts to support attentional selection shifts beyond the high acuity range are implemented on shared computational circuits and underlying substrate but distinguished by their own functions for visual selection. The shifting patterns of the presented model have not been compared with human behaviour as few specific findings are available. Nevertheless, the framework is inspired by recent psychophysical research on multiple attentional selectivity and their relationships with eye movements. Its performance has been demonstrated in natural scenes and shows the ability to effectively reduce the shift times and search errors to select useful objects, regions, and structured groups, and the ability to flexibly select visual-objects whether located in the foveal field or in visual periphery. It borrowed automatic segmentation and partition methods from other works to create dynamic visual-objects. It is very interesting to investigate in the future whether unitary attention can be implemented without specific perceptual grouping or with very rough perceptual grouping, and whether these two procedures can be modelled to facilitate each other.

## Acknowledgements

This paper and Y. Sun are supported by National Natural Science of Foundation of China (NSFC No. 60775019).

## References


[1] G. Backer, B. Mertsching, and M. Bollmann, "Data- and model-driven gaze control for an active vision system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1): 1-15, 2002.

[2] G. Bonmassar, E. L. Schwartz, "Space-Variant Fourier Analysis: The Exponential Chirp Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10), pp. 1080-1089, 1997.

[3] C. D. Chambers and J. B. Mattingley, "Neurodisruption of selective attention: insights and implications," *Trends in Cognitive Sciences*, 9(11), pp. 542-550, 2005.

[4] L. Craighero and G. Rizzolatti, "The premotor theory of attention," In: *Neurobiology of attention*, L. Itti, G. Rees, and JK Tsotsos (Eds), New York, Elsevier, pp. 181-186, 2005.

[5] B. Draper and A. Lionelle, "Evaluation of Selective Attention Under Similarity Transforms," *Workshop on Performance and Attention in Computer Vision*, Graz, Austria, pp. 31-38, April 3, 2003.

[6] J. Duncan, "Converging levels of analysis in the cognitive neuroscience of visual attention," *Phil. Trans. R. Soc. Lond. B.*, 353, pp. 1307-1317, 1998.

[7] H. M. Gomes, R. B. Fisher, *Model Learning in Iconic Vision*, PhD Thesis, School of Informatics, The University of Edinburgh, 2002.

[8] Y. Haxhimusa and W. G. Kropatsch, "Hierarchical image partitioning with dual graph contraction," in *Proc. of 25th DAGM Symposium LNCS*, B. Milaelis and G. Krell (Eds.), 2781, pp. 338-345, 2003.

[9] J. E. Hoffman, "Visual attention and eye movements," In H. Pashler (Ed.), *Attention*, Psychology Press, pp. 119-154, 1998.

[10] T. K. Horiuchi, T. G. Morris, C. Koch and S. P. DeWeerth, "Analog VLSI Circuits for attention-based visual tracking," The Neural Information Processing Conference, Denver CO, pp. 706-712, December, 1996.

[11] T.S. Horowitz, E. M. Fine, D. E. Fencsik, S. Yurgenson, and J. M. Wolfe, "Fixational eye movements are not an index of covert attention," *Psychological Science*, 18(4), pp. 356-363, 2007.

[12] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pp. 1254-1259, 1998.

[13] http://ilab.usc.edu/bu/javaDemo/index.html.

[14] CH Juan, SM Shorter-Jacobi, and JD Schall, "Dissociation of spatial attention and saccade preparation," Proceedings of the National Academy of Sciences of the United States of America, 101, pp. 15541-15544, 2004.

[15] T. A. Kelley, J. T. Serences, B. Giesbrecht, and S. Yantis, "Cortical Mechanisms for Shifting and Holding Visuospatial Attention," *Cerebral Cortex*, January 1, 18(1), pp. 114 - 125, 2008.

[16] E. Kowler, "Attention and eye movements," In *The New Encyclopedia of Neuroscience*, Volume editor: R. Krauzlis, Elsevier, Amsterdam, in press.

[17] D. LaBerge, *Attentional Processing: the Brain's Art of Mindfulness*, Harvard University Press, 1995.

[18] K. Lee; H. Buxton, J. Feng, "Cue-guided search: a computational model of selective attention," *IEEE transactions on neural networks*, 16, pp. 910-924, 2005.

[19] R. M. McPeek, V. Maljkovic and K. Nakayama, "Saccades require focal attention and are facilitated by a short-term memory system," *Vision Research*, 39, pp. 1555-1566, 1999.

[20] F. Orabona, G. Metta, G. Sandini, "Object-based Visual Attention: a Model for a Behaving Robot," In 3rd International Workshop on Attention and Performance in Computational Vision within CVPR, San Diego, CA, USA. June 25, 2005.

[21] H. Pashler, *The Psychology of Attention*, Cambridge, MA: MIT Press, 1998.

[22] M. E. Posner, Y. Cohen, and R. D. Rafal, "Neural systems control of spatial orienting," *Phil. Trans. R. Soc. Lond. B*, 298, pp. 187-198, 1982.

[23] Z. W. Pylyshyn, "Visual indexes, preconceptual objects, and situated vision," *Cognition*, 80, pp. 127-158, 2001.

[24] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, 7, pp. 17-42, 2000.

[25] *http://www.caip.rutgers.edu/riul/*.

[26] G. Rizzolatti, L. Riggio, BM Sheliga, "Space and selective attention," In *Attention and performance XV*, M. Moscovitch and C. Umilta (Ed.), Cambridge, MA: MIT Press, pp. 232C265, 1994.

[27] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, 80, pp. 1-46, 2001.

[28] G. Sela and M. D. Levine, "Real-time attention for robotic vision," *Real-Time Imaging*, 3, pp. 173-194, 1997.

[29] J. T. Serences, T. Liu, and S. Yantis, "Parietal mechanisms of switching and maintaining attention to locations, objects, and features," In L. Itti, G. Rees and J. Tsotsos (Eds.), *Neurobiology of Attention*, New York: Academic Press, pp. 35-41, 2005.

[30] Y. Sun and R. Fisher, "Object-Based attention for computer vision," *Artificial Intelligence*, 146 (1), pp. 77-123, 2003.

[31] Y. Sun, *Hierarchical Object-Based Visual Attention for Machine Vision*, PhD Thesis, School of Informatics, The University of Edinburgh, 2003.

[32] K. G. Thompson, K. L. Biscoe, and T. R. Sato, "Neuronal basis of covert spatial attention in the frontal eye field," *The Journal of Neuroscience*, 25, pp. 9479-9487, 2005.

[33] S. Tipper, B. Weaver, L. Jerreat, and A. Burak, "Object-based and environment-based inhibition of return of visual attention," *Journal of Experimental Psychology: Human Perception and Performance*, 20, pp. 478-499, 1994.

[34] J. K. Tsotsos, Y. Liu, J. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou, "Attending to visual motion," *Computer Vision and Image Understanding*, 100, pp. 3-40, 2005.

[35] R. D. Wright and C. M. Richard, "Inhibition-of-return at multiple locations in visual space," *Canadian Journal of Experimental Psychology*, 50(3), pp. 324-327, 1996.

[36] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, 19, pp. 1395-1407, 2006.

[37] A. Zaharescu, A. Rothenstein, J. K. Tsotsos, "Towards a Biologically Plausible Active Visual Search Model," Proc. ECCV WAPCV2004, Prague, May 15, pp. 133-147, 2004.